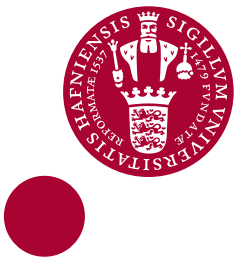

What's in the Box?

The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration

Olsen, Henrik Palmer; Slosser, Jacob Livingston; Hildebrandt, Thomas Troels; Wiesener, Cornelius

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration

Henrik Palmer Olsen, Jacob Livingston Slosser, Thomas Troels Hildebrandt & Cornelius Wiesener

University of Copenhagen Faculty of Law
Legal Studies Research Paper Series, paper no. 2019-84

FACULTY OF LAW
UNIVERSITY OF COPENHAGEN

iCourts Working Paper Series, No. 162, 2019

What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration

Henrik Palmer Olsen, Jacob Livingston Slosser, Thomas Troels Hildebrandt, and Cornelius Wiesener

iCourts - The Danish National Research Foundation's
Centre of Excellence for International Courts

June 2019

Abstract:
Every day, millions of administrative transactions take place. Insurance policies, credit appraisals, permit and welfare applications, to name a few, are created, invoked, and assessed. Though often treated as banalities of modern life, these transactions often carry significant importance. To the extent that such decisions are embodied in a governmental, administrative process, they must meet the requirements set out in administrative law, one of which being the requirement of explainability. Increasingly, many of these tasks are being fully or semi-automated through algorithmic decision making (ADM) systems. Fearing the opaqueness of the dreaded black box of these ADM systems, countless ethical guidelines have been produced for combatting the lack of computational transparency. Rather than adding yet another ethical framework to an already overcrowded ethics-based literature, we focus on a concrete legal approach, and ask: what does explainability actually require? Using a comparative approach, we investigate the extent to which such decisions may be made using computational tools and under what rubric their compatibility with the legal requirement of explainability can be examined. We assess what explainability actually demands with regard to both human and computer-aided decision-making and which recent legislative trends, if any, can be observed. We also critique the field's unwillingness to apply the standard of explainability already enshrined in administrative law: the human standard. Finally, we introduce what we call the "administrative Turing test" which could be used to continually validate and strengthen AI-supported decision-making. With this approach, we provide a benchmark of explainability on which future applications of algorithmic decision-making can be measured in a broader European context, without creating an undue burden on its implementation.

KEYWORDS: explainability, algorithmic decision making, administrative law, artificial intelligence, black box

Henrik Palmer Olsen, Professor of Jurisprudence, Faculty of Law, Danish National Research Foundation's Centre of Excellence for International Courts (iCourts), University of Copenhagen; henrik.palmer.olsen@jur.ku.dk

Jacob Livingston Slosser, Carlsberg Postdoctoral Research Fellow, Faculty of Law, Danish National Research Foundation's Centre of Excellence for International Courts (iCourts), University of Copenhagen; jacob.slosser@jur.ku.dk

Thomas Troels Hildebrandt, Professor in Software Engineering, Software, Data, People & Society Section, Department of Computer Science, University of Copenhagen, Denmark; hilde@di.ku.dk

Cornelius Wiesener, Postdoctoral Research Fellow, Faculty of Law, Danish National Research Foundation's Centre of Excellence for International Courts (iCourts), University of Copenhagen; cornelius.wiesener@jur.ku.dk

This research is funded by the Danish National Research Foundation Grant no. DNRF105.

iCourts - Centre of Excellence for International Courts - focuses on the ever-growing role of international courts, their place in a globalizing legal order, and their impact on politics and society at large. To understand these crucial and contemporary interplays of law, politics, and society, iCourts hosts a set of deeply integrated interdisciplinary research projects on the causes and consequences of the proliferation of international courts.

iCourts opened in March 2012. The centre is funded by a large grant from the Danish National Research Foundation (for the period 2012-22).

What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration

Henrik Palmer Olsen, Jacob Livingston Slosser, Thomas Troels Hildebrandt, and Cornelius Wiesener

1. Introduction

As the quality of AI improves, it is increasingly applied to support decision-making processes, including in public administration. This has many potential advantages: faster response time, better cost effectiveness, better quality, etc. At the same time, implementing AI in public administration also raises a number of concerns: bias in the decision-making process, lack of transparency, elimination of human discretion, among others.[1] Often, these concerns are raised to a level that obscures the legal remedies that exist to curb those fears, and unduly delays the implementation of efficient systems. The fears raised by the administrative use of AI systems are threefold.[2] First, is the loss of control over systems and thus, a clear link to responsibility when decisions are taken.[3] In a discretionary system, someone must be held responsible for those decisions and be able to give reasons for them. There is a legitimate fear that in a black box system used to produce a decision, even when used in coordination with a human counterpart or oversight, creates a system that lacks responsibility. This is the fear of the rubber stamp: that, even if a human is in the loop, the deference given to the machine is so much that it creates a vacancy of accountability for the decision.[4] The second fear of algorithmic decision making (ADM) systems is a loss in human dignity. If legal processes are replaced with algorithms, there is a fear that humans will be reduced to mere "cogs in the machine". Rather than being in a relationship with other humans to which you can explain your situation, you will be reduced to a digital representation of a sum of data. Since machines cannot reproduce the whole context of the human and social world, but only represent specific limited data about a human (say age, marital status, residence, income, etc.), the machine cannot *understand* you. Removing this ability to understand and to communicate freely with another human can easily lead to alienation and a loss of human dignity. Lastly, there is the well documented fear of 'bad' data being used to make decisions that are false and discriminatory. These decisions range from the use

[1] See among various others, Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2018); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2017).

[2] Lee A Bygrave, 'Article 22 Automated Individual Decision-Making, Including Profiling', *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2019) <https://works.bepress.com/christopher-kuner/2/download>.

[3] A related, but more legal technical problem in regards to the introduction of AI public administration is the question of *when* exactly a decision is made. Associated to this is also the problem of *delegation*. If a private IT developer designs a decision-system for a specific group of public decisions, does this mean that those decisions have been delegated from the public administration to the IT developer? We shall not pursue these questions in this paper.

[4] Elin Wihlborg, Hannu Larsson and Karin Hedstrom, '"The Computer Says No!" -- A Case Study on Automated Decision-Making in Public Authorities', *2016 49th Hawaii International Conference on System Sciences (HICSS)* (IEEE 2016) <http://ieeexplore.ieee.org/document/7427547/>.

of false profiling to self-reinforcing feedback loops that can be a significant breach of law if not just societal norms.[5]

While we accept that these fears are not unsubstantiated, they needn't prevent existing legal remedies from being acknowledged and used. Legal remedies should be used rather than the more cursory and sometimes naive reach towards general guidelines or grand and ambiguous ethical press releases, that are not binding, not likely to be followed, and do nothing to solve the real problems they hope to address. In order to gain the advantages of AI-supported decision-making, these concerns must be met by indicating how AI can be implemented in public administration without undermining the qualities associated with contemporary administrative procedures. In this paper, we focus on how AI-supported administrative decision-making can be introduced in such a way that it meets the explainability requirement in administrative law: administrative decisions addressed to citizens must be supplied with a relevant explanation for that decision.[6]

The paper examines the explainability requirement as follows: first, we outline how explainability should be understood as *legal* explainability rather than *causal* explainability (section 2), dismissing the idea that transparency in AI-supported decision-making necessarily implies mathematical transparency.[7] To illustrate the single legal problem that exists under a regime of explainability, we apply these rules to a scenario based on real world casework that exists as both human-only and ADM systems.[8] This scenario results in three models of when explanation is given: a pure model (only human decision makers); a hybrid model (with some combination of human and ADM); and, a fully automated model. Each of these models would go through a decision phase and, if needed, an appeal phase, both of which could be described by one of the models. Unlike some calls for recourse for 'human in the loop' models, we argue that the simple existence of human intervention does not address (and is subsidiary to) the stronger requirement for legal explainability. To give each model a tangible grounding, we consider each under the scenario of an administrative decision regarding the Danish law on the requirement on municipalities to provide a compensation for loss of earnings to parents who provide care to a child with permanent reduced physical or mental functioning (in particular whether an illness would be considered "serious, chronic or long-term").[9]

---

[5] For an overview of the social dangers associated with AI more generally, see Iyad Rahwan and others, 'Machine Behaviour' (2019) 568 *Nature* 477.

[6] By explanation, we mean here that the administrative agency gives reasons that support its decision. In this paper, we use the term explainability in this sense. This is different from explainability used in relation to the so-called "black box problem", Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1 *Nature Machine Intelligence* 206. As we explain below, we think the quest for black box explainability (which we call mathematical transparency) should give way to explainability in the legal sense (giving grounds for decisions). We take this to be in line with Rudin's call for interpretability in high stakes decisions.

[7] See the debate outlined in: Brent Daniel Mittelstadt et al, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) *Big Data & Society* 6–7.

[8] See, Ecoknow project: https://ecoknow.org/about/

[9] "Persons maintaining a child under 18 in the home whose physical or mental function is substantially and permanently impaired, or who is suffering from serious, chronic or long-term illness. Compensation shall be subject to the condition that the child is cared for at home as a necessary consequence of the impaired function, and that it is most expedient for the mother or father to care for the child." § 42 (1) of the Danish Consolidation Act on Social Services, available at: http://english.sm.dk/media/14900/consolidation-act-on-social-services.pdf. For a review of the legal practice in municipalities, see: Ankestyrelsen, 'Ankestyrelsens Praksisundersøgelse Om Tabt Arbejdsfortjeneste

We then outline the different types of decision-making systems: rule-based, machine-learning-based, and hybrid approaches (section 3). These different approaches create a range of different applications that have their own unique obstructions in regards to legibility, interpretation and transparency. This range of applications shows that simply adding a human in the loop at an (arguably) indiscriminate point along this spectrum creates an undue strain on AI systems in public administration and ignores the legal remedies that already exist. We emphasize that this does not mean that we are opposed to advancing algorithmic transparency in any kind or form. Nor does this mean that this new form of decision-making could not advance better explanations and enhance legal certainty by allowing deeper insights into the relationship between facts and legal arguments in administrative decisions. On the contrary, we generally support the push for algorithmic transparency. However, we object to the argument that AI-supported decision-making cannot be introduced (i.e. that it will be illegal or ethically contentious to do so) in public administration, unless it is algorithmically transparent. The introduction of AI-supported decision-making should not be prevented by new and stricter requirements specifically aimed at such decision-making.

Next (section 4), we look at what the *explainability* requirement means. We do this by breaking up the requirement into a number of smaller elements and illustrate this with examples from various national (Denmark, Germany, France, and the UK) and regional legal systems (EU law and the European Convention of Human Rights). Given the wide range of legal approaches and the firm foundation of the duty to give reasons, we argue that the requirements attached to the existing standards of explainability are well-tested, adequate, and sufficient to protect the underlying values behind those standards. AI-supported decisions can and should be held accountable under those existing legal standards and that any arguments about the minimum requirements in regards to AI-supported decision-making should be set at the same threshold of explainability set for purely human-based decisions. If the arguments set a higher standard for explainability or transparency in AI-supported than in solely human decision-making, then the arguments will not be valid, because they will introduce a legal standard different from that which exists in the current law. Rather than introducing new legal requirements, a more dynamic communicative process aimed at citizen engagement with the algorithmic processes employed by the administrative agency in question will be more suitable to advancing the overall legitimacy of using AI technology in public administration. As an example model of what this process might look like, we introduce our novel solution, what we call the "administrative Turing test" (section 5). This test could be used to *continually validate and strengthen* AI-supported decision-making. As the name indicates, it relies on comparing solely human and algorithmic decisions, and only allows the latter when an administrative caseworker cannot immediately tell the difference between the two. The administrative Turing test is an instrument to ensure that the analogue explainability test is met in practice. Using this test in administrative decision-making systems is aimed at ensuring the continuous sensitivity of law to its context (i.e. avoiding unwanted rigidity in the application of law) and advancing human trust in the AI-generated

Efter Servicelovens § 42 (National Board of Appeal's Study on Lost Earnings According to Section 42 of the Service Act)' (2017) https://ast.dk/publikationer/ankestyrelsens-praksisundersogelse-om-tabt-arbejdsfortjeneste-efter-servicelovens-ss-42.

output. Implementing this test also advances what – according to some of the latest research – is the best way to use AI for legal purposes, namely in a set up that relies on AI and human collaboration.[10]

2. Critique of Bifurcating Explainability – Human v Machine

The explainability requirement (or duty to give reasons) serves a number of functions in public administrative law. First, it enhances the legitimacy of the decision in question and of public administrative decision-making in general. By providing reasons, the decision is legitimized by showing *how* the decision is lawful. Secondly, requiring explainability enhances reflexivity in the administrative process towards decision-making, thereby improving the quality of those decisions by more carefully testing the facts of the case against the legal requirements that apply to the case at hand. Finally, the explainability requirement eases the review process should the decision be appealed to a higher administrative body or the courts.[11]

The explainability requirement for administrative decisions can be found, in one guise or another, in most legal systems. In Europe, it is often referred to as the "duty to give reasons", i.e. a positive obligation to provide an explanation ("begrundelse" in Danish, "Begründung" in German and "motivation" in French). The explainability requirement is closely linked to the right to legal remedies. In fact, its emergence throughout history was driven by the need to enable the citizen affected by an administrative decision to effectively challenge it before a court of law.[12] This, in turn, required the provision of sufficient reasons for the decision in question: both towards the citizen, who as the immediate recipient should be given a chance to understand the main reasoning behind the decision, and the judges, who will be charged with examining the legality of the decision in the event of a legal challenge. The duty to give reasons has become a self-standing legal requirement, serving a multitude of other functions beyond ensuring effective legal remedies, such as: ensuring better clarification, consistency and documentation of the decisions, self-control of the decision-makers, internal and external control of the administration as a whole, as well as general democratic acceptance and transparency.[13]

---

[10] See, Saul Levmore and Frank Fagan, 'The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion' (forthcoming 2019) 93 *Southern California Law Review* available at: https://papers.ssrn.com/abstract=3362563.

[11] See, Carol Harlow and Richard Rawlings, 'Proceduralism and Automation: Challenges to the Values of Administrative Law' in E Fisher, J King and A Young (eds), *The Foundations and Future of Public Law (in honour of Paul Craig)* (OUP Oxford 2019) https://papers.ssrn.com/abstract=3334783, 12–13 (SSRN version).

[12] Uwe Kischel, *Die Begründung: Zur Erläuterung Staatlicher Entscheidungen Gegenüber Dem Bürger*, vol 94 (Mohr Siebeck 2003) 32–34.

[13] Franz-Joseph Peine and Thorsten Siegel, *Allgemeines Verwaltungsrecht* (2018, 12th ed., C.F. Müller), 160, mn. 513; Schweickhardt, Vondung, Zimmermann-Kreher (eds), Allgemeines Verwaltungsrecht (2018, 10th ed., Kohlhammer, Stuttgart), 586-88; Uwe Kischel, 2003, 40-65; H. C. H. Hofmann, G. C. Rowe, A. H. Türk, *Administrative Law and Policy of the European Union*, (Oxford University Press 2011), 200–202; CJEU, *Bamba v Council, Judgment*, 15 November 2012, Case C-417 / 11, para. 49; N. Songolo, La motivation des actes administratifs, 2011, www.village-justice.com/articles/motivation-actes-administratifs,10849.html; J.-L. Autin, La motivation des actes administratifs unilatéraux, entre tradition nationale et évolution des droits européens "RFDA" 2011, no. 137-38, 85-99. The transparency aspect is further strengthened by the more recent evolution of the freedom of information. While often conflated with the explainability requirement, both denote two separate legal concepts with different procedural frameworks. In this paper, we will primarily focus on the explainability requirement.

It seems to follow implicitly from the explainability requirement that what counts as explanation must refer to the law that undergirds the decision (and the facts that are relevant to the case by virtue of the law). The explainability requirement should be understood in terms of the law that regulates the administrative body's decision in the case before it. It is not a requirement that *any* kind of explanation must be given but rather a *specific kind* of explanation. This observation has a bearing on the kind of explainability that may be required for administrative decision-making relying on algorithmic information analysis as part of the process towards reaching a decision.

Take, for instance, our example of Parent A. An administrative body issues a decision to Parent A in the form of a rejection explaining that the illness the child suffers from does not qualify as *serious* within the meaning of the statute. The constituents of this explanation, while varying in differing jurisdictions, would generally cover a reference to the child's disease and the qualifying components of the category of *serious illness* being applied. This could be anywhere from a checklist system (one might say a human algorithm) or reference to a list of diseases that qualify and an explanation of the differences between the applicant disease and those categorised as applicable under the statute. Perhaps it might also include alternatives for consideration of a positive decision (a GP's note, etc.). In general it would explain:

1. the legislative grounds on which the decision rests,
2. the salient facts of the case, and
3. the most important connection points between them, i.e. the discretionary or interpretive elements that are attributed weight in the decision-making process.[14]

It is against this background that the threshold for explainability should be understood.

In a purely human system, at no point would the administrative body be required to describe the neurological activity of the caseworkers that have been involved in making the decision in the case. Nor would they be required to provide a psychological profile and biography of the administrator involved in making the decision giving a history of the vetting and training of the individuals involved, their educational backgrounds, to account for all the inputs that may have been explicitly or implicitly used to consider the application. Nor would the human decision making system in general, in all its biological, social, and psychological complexity, be legally required to be outlined within the explanation.

Enter the machine.

When the same process involve an ADM system, must the explanation open up the opaqueness of its mathematical weighting? Must it provide a technical profile of all the inputs into the system? In the case of a hybrid system with a human in the loop, must the administrators set out – in detail – the

[14] Making sure that the connection relies on "clean" data is a separate issue that we do not touch on in this paper. For discussion of this issue in regards to AI supported law enforcement, see: Rashida Richardson, Jason Schultz and Kate Crawford, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice' [2019] *New York University Law Review Online*, Forthcoming.

electronic circuits that connect the computer keyboard to the computer hard drive and the computer code behind the text-processing program used? Must it describe the interaction between the neurological activity of the caseworker's brain and the manipulation of keyboard tabs leading to the text being printed out, first on a screen, then on paper, and finally sent to the citizen as an explanation of how the decision was made?

Obviously, requiring such high levels of explanation is both insufficient and superfluous. Even though it may be empirically fully accurate, it does not meet the requirement of *legal* explanation. It gives an explanation – but it does not give to the citizen the explanation he or she is looking for. The problem in this example is that the explanation provided does not connect the decision to its legal basis. It is, in other words, not possible to see the *legal reasoning* leading from the facts of the case to the legal decision. The reasons that make information about the neurological processes inside the brains of caseworkers, or their biographical histories irrelevant to the explainability requirement are the same that make information about the algorithmic calculus (sometimes referred to as "transparency") in an administrative support system similarly irrelevant. This is not as controversial of a position as it might seem on first glance.

We would like to emphasize that ADM in public administration is a phenomenon that comes in a wide range of formats: from the use of automatic information processing for use as one part of basic administrative decisions (already in use in tax administration in many countries, where some, but not all income is automatically processed in the calculation of a citizen's annual tax duties), over semi-automated decision-making, used for example in predictive policing and other types of profiling systems for example control systems for business regulation, to fully automated decision-making that uses AI to link information about facts to legal rules via machine learning. There is, in other words, a wide spectrum of ways in which AI technology can be used to support administrative decision-making.

While fully automated models have attracted a lot of attention,[15] we examine a hybrid model for AI-supported decision-making that applies as a collaboration between AI and human intelligence in the decision-making process. This focus, we think, is the most interesting because fully automated models can only (so far) be applied to simple forms of legal casework. The fully automated system as it is further off will be dealt with in subsequent studies. In the meantime, the frontline in research we claim, is precisely in the construction of collaborative models that enhance efficiency and quality in administration when compared to the fully human model. Only such models are likely to have an impact on legal decision-making in public administration in the near future.[16] We submit that the existing legal standards in relation to explainability should be preserved and neither strengthened nor loosened *as a result of introducing AI as part of the case-handling process in public administration*. We argue for an ADM procedure that will result in decisions reaching the same level of explainability as is demanded under existing law from purely human decision-making. The system we envisage,

[15] Perhaps most famous is O'Neil (n 2), but the debate on Technological Singularity has attracted a lot of attention, see for an overview: Murray Shanahan, *The Technological Singularity* (MIT Press 2015).

[16] See Saul Levmore and Frank Fagan, 'The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion' (2019) 93 *Southern California Law Review* available at: https://papers.ssrn.com/abstract=3362563.

therefore is based on the idea that AI-support comes in the form of an algorithmic system that provides drafts of decisions to human caseworkers by highlighting the three requirements set out above. In this process, AI-support is applied at the drafting stage. The algorithmic system then is used only to prepare the human-made decision by providing an AI-generated decision proposal. Since the explainability requirement relates to legal reasons and the facts of the case a lack of transparency in regards to the algorithmic calculus used to draft a decision in the case cannot in and of itself invalidate the decision.[17]

An obvious rebuttal stems from the fear that the human becomes a simple rubber stamp to the computer-generated decision.[18] This is problematic if it results in a false sense of confidence in the decision. This will be the case only if the AI-generated drafts that are being rubber-stamped (i.e. not made subject to substantive critical review by the human) are *of a lower legal quality* than what could have been provided via purely human decision-making. Referring back to our introductory remarks, the standard for decision-making quality that has to be met (but not necessarily improved however desirable this may be) is the human standard.[19]

Retaining a human standard for explainability, rather than introducing a new standard devised specifically for AI-supported decision-making, has the extra advantage that the administrative agency remains fully responsible for the decision. With the requirement that decisions be legible follows that they must be meaningful.[20] From this also follows that the administrative agency issuing the decision can be queried about the decision in ordinary language. This then assures that the rationale behind the explainability requirement is respected even if the decision has been arrived at through some algorithmic calculation that is not transparent.

If the analogy is apt in comparing algorithmic mathematics to human neurology or biography, then requiring algorithmic transparency in legal decisions that rely on AI-supported decision-making would be to fail to address the explainability requirement at the right level. Much in line with Rahwan et al, who argue for a new field of research – the study of machine behaviour akin to human behavioural research[21] – we argue that the inner workings of an algorithm is not what is in need of explanation, but rather, the human interaction with the *output* of the algorithm. AI-supported decision-making should not be required to have a more finely granulated level of explainability than human decision-making before it can be put to use in public administration. AI-supported decision-

[17] The procedure can be likened to that of a professional translator who uses Google Translate to create a draft translation, which the translator then works over to create the final translation. The fact that Google Translate is based on a secret algorithm, does make the final translation provided by the translator a less valuable translation than if the translator had written up the translation from scratch. What matters is the quality of the final product, not how it was produced.

[18] This issue was raised in discussion over the case *State vs. Loomis* (Wisconsin Supreme Court), which concerned the use of a non-transparent algorithm in making a judicial decision about sentencing. The case is discussed in 130 *Harvard Law Review*, 1530 (2019). See https://harvardlawreview.org/2017/03/state-v-loomis/

[19] A consequence of this is that AI supported decisions must be legible. This point has been made already in regards to the GDPR, see Gianclaudio Malgieri and Giovanni Comandé, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 243.

[20] ibid.

[21] See Rahwan and others (n 6).

making has much potential value and the introduction of this technology should not be prevented by raising the bar of explainability to a higher level than what exists today for human decision-making.

Parent A is no more satisfied by an explanation that gives reasons to the inner workings of an ADM system than they would be with what a biologist might have to say about a caseworker's brain. To see this, let us explore what this kind of explanation would provide.

3. Administrative Decision-Making via Machine Models

Machine approaches to this kind of ADM are about classifying members of a domain of interest into different types. In decision support for case management in public administration, the domain of interest is *cases* (i.e. the data describing the case), and the types are the possible outcomes of the cases. However, a case will normally require a number of sub-decisions. In our example, the overall outcome is whether or not to pay compensation for loss of earnings to a parent. In this section alone, there are a number of sub-decisions to be made before one can decide to pay compensation. To make these decisions computationally, there are many approaches. At the highest level, we speak about: rule-based, machine-learning-based, and hybrid approaches (combining the two).

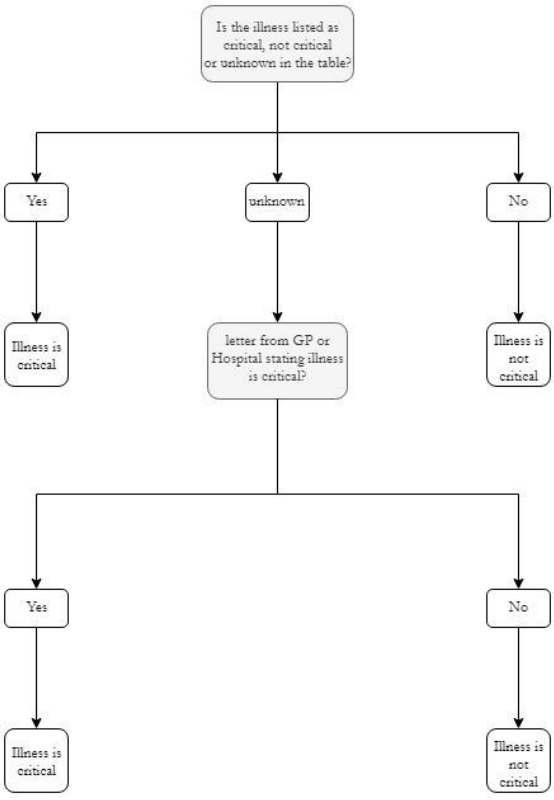

Figure 1

A rule-based approach refers to what was introduced as *expert systems* in the 1970s. In a rule-based expert system, rules are given by experts. For instance, a rule for deciding when an illness considered to be *serious, chronic or long-term illness* may be given as a table of illnesses. Such a table will need to be revised over time, e.g. when new types of illnesses or treatments are discovered. A rule may also be that a statement from the general practitioner (GP) or hospital that the illness is serious, chronic or long-term will serve as proof of fact. Finally, a third rule may be used to combine the two rules, stating that if the illness is not on the list given in the first rule, then the second rule and the statement from the GP or hospital serves as proof of fact.

In machine-learning-based approaches, the rules are not programmed by experts, but computed by an algorithm from a set of example members of the domain of interest, referred to as the training set. In so-called supervised learning, the training set is assumed to be *a priori* correctly classified in types, and then an algorithm fed by the examples computes a set of rules that can reconstruct this classification to a high degree of precision. A very naïve algorithm would use the training set as the "rule" and say "approve" if exactly the same case has been seen

before and been approved and otherwise say no. In situations where the number of variations of cases is high or perhaps even infinite, this will inevitably give rise to many false rejections, i.e. rejections of cases that should have been approved but simply were not seen before in the example data. Instead, algorithms generally use a finite number of features (e.g. a given name of an illness, a statement from the GP, a statement of the hospital) and then determine rules for how these features should influence the decision.

There are hundreds of different algorithms, referred to as *classifiers*, prescribing how to best determine features from the example data and how they determine the outcome. The algorithms for classifiers can themselves be classified in different families developed in different areas of computer science and mathematics. Among the best known are: decision trees, random forests, rule-based (e.g. expert systems as exemplified above) and neural networks.

In the decision tree approach (*fig 1*), the algorithm builds a single decision tree, starting with a classifying question in the root, and answers to the question leading to a sub-node with a new question or to a leaf with an answer. For Parent A, the root could be the question “Is the illness listed as critical, not critical or unknown in the table?”. The branch with the answer “Yes” could then lead to a leaf with the answer “illness is critical”. The branch with the answer “No” could lead to the leaf with the answer “illness is not critical”. Finally, the branch with the answer “Unknown” could leave to a sub node with the question “letter from GP or hospital stating illness is critical?” with two outgoing branches, labelled "yes" and "no" leading to answer leaves labelled respectively "illness is critical" and "illness is not critical". The precision of a decision tree for a binary classification is measured according to:

- true positives (answers for cases in the training set that are correctly classified as critical illness),
- false positives (answers for cases in the training set that are classified as critical illness by the algorithm, but are not *a priori* classified as such),
- true negatives (answers for cases in the training set that are correctly classified as not being critical illness by the algorithm, and
- false negatives (answers for cases in the training set that are classified as not being critical illness by the algorithm, but are *a priori* classified as being critical illness).

The precision is highly dependent on the choice of features (questions to ask), and it may be part of the algorithm to detect "good" features. A feature of the case that may be good at determining the outcome may, however, not be good for other reasons, i.e. it may be that the algorithm in the training set detects features such as the gender or the income of the citizen, or even a characteristic way of using punctuation in the notes describing the case, which are not (legally) valid features on which to base the decision. Even for valid features, there may be several competing choices of features and the order in which to ask the questions. As has been seen in numerous cases, the invalidity of these choices is a major concern for algorithmic systems, and the main impetus for calls for explainability.

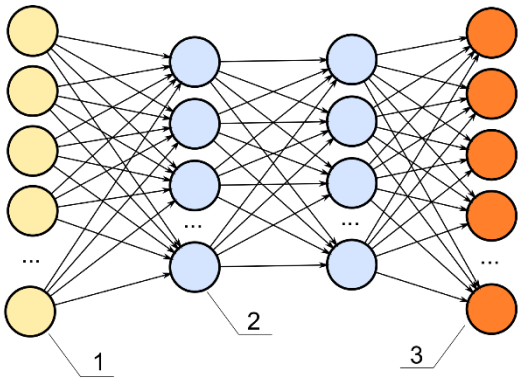

Figure 2
Zufzzi, 2010,
https://commons.wikimedia.org/wiki/File:Neural_network_bottleneck_architecture.svg

Random forest approaches construct an ensemble of decision trees (i.e. a forest) by selecting the features (questions to ask) randomly for each tree. The final outcome will then be based on the outcome of all the decision trees, e.g. by taking the majority vote. This can be compared to having a number of case workers, each focusing on different features.[22] A study from 2014 evaluated 179 different classifiers and found that three of the top five belonged to the random forest family.[23] The random forest family is a generalization of the decision tree approach. A hybrid approach would have decision trees (or forests) where a rule for branching in a node is either learned from the training data or given by an expert. This would also describe the situation where an expert can overrule a learned rule for branching in a tree for a concrete case.

This classification task gets more complex to explain to Parent A as you add in deep learning to the process, such as the oft-maligned neural network approach (of which there are many different variants). In general, and extremely simplified, a neural network consists of multiple 'neurons' each representing a potential to activate or not according to a threshold value. Each set of neurons can be classified in one of three layers as illustrated in fig 2: input (1), hidden (2), and output (3). In our scenario, the nodes in the input layer correspond to the features (.e.g data about individuals, contexts, and situations) where a decision was made about a 'serious' illness. This first layer of neurons in connected to the next layer through weighted connections (where a value is added to the connection). This next neuron gets its value from the weighted value of the incoming connections and the neuron(s) that come before it. This value then determines whether the neuron then sends its value forward to the next layer. Depending on the incoming values, the hidden layers (of which there can be many) then determines whether the value sent forward is enough to fire the output layer which will give the determination of a classification. The system is structured in a way that the neurons can replicate patterns that might not be easy to spot or available for a simple rule based system. As the system

[22] We make this last point in response to a possible fear against using AI for decision-making in public administration. It could be argued that in a human only bureaucracy, decisions are discussed in groups between case-worker, for example as a way of advancing best practice standards, or when countering cases that are novel and/or raise new principle issues. Such group discussions can interpreted as a bureaucratic process which relies on collective reasoning for a best result. Introducing an AI support structure for decision-making could be seen as a way of centralizing the decision-making process (or parts thereof), thereby effectively eliminating the collective reasoning process. As we emphasize towards the end of this paper, we ultimately argue for a hybrid approach (human-AI collaboration), but at this point we emphasize that the random forest approach to machine learning could perhaps mitigate the feared centralizing effect.

[23] Manuel Fernández-Delgado and others, 'Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?' (2014) 15 *J. Mach. Learn. Res*. 3133.

learns, the weights and activation values are refined to become more precise, where precision should be understood as described above.[24]

Though there are efforts to make more transparent systems, the idea that explaining the mechanism by which a system made its decision doesn't satisfy the legal requirements stemming from explainability. There is an inherent irony of asking transparency from a computational neural network loosely modelled on the structure of the brain and not asking it of its opaque biological counterpart. Furthermore, given the numerous combinations of approaches and models of decision-making, there will be a spectrum of scenarios including some combination of human and automated decision-making. Separate legal requirements for each point along this spectrum, or a threshold that is arbitrarily placed given the presence (or not) of a human, or the presence (or not) of a machine, seems an unnecessary burden to administrative practice. It is a superfluous addition to legal explainability as already enshrined through various legal instruments.

4. Explainability in Comparison: Legal Doctrine

In this section, we examine the relevant legal framework in Denmark, Germany, France and the United Kingdom. While limited in scope, this case selection includes a variety of different legal cultures across Europe as well as different stages of developing digitalised administrations (i.e. both front-runners and late-comers in that process). Complementing this picture, we also consider the additional legal standards arising from EU law and European human rights law. We outline generally, what explainability actually demands. By outlining the threshold of explainability as it already exists, we highlight the lack of a requirement for additional legal principles for the implementation of ADM systems in administrative law. The principle of explainability, as it stands, is a version of algorithmic transparency that is sufficient to assuage the fears surrounding ADM.

a. Denmark

The Danish Act on Public Administration came into force in 1987. It sets out a general framework for the operation of public administration and the rights of citizens in regards to administrative decision-making that affects their rights and interests.[25] The act was later supplemented by the right to information act, which gives every citizen a right (with certain exceptions) to obtain information from public bodies.[26]

[24] This refinement can often be fully opaque as to what is 'decisive' as an input in the system, and that some inputs are reproductions of 'bad data'. This is certainly true, but only different in degree and not kind from a 'bad data' problem in a pure human system. The problem of bad data is a broader problem of classification that exists outside of the introduction of ADM. Classification, in itself, has long been known as an exercise that is inherently an act of discrimination that too often is premised on distancing the already marginalised, and reinforcing stubborn biases. See, Geoffrey C Bowker and Susan Leigh Star, *Sorting Things out: Classification and Its Consequences* (MIT press 2000).

[25] See, more generally: Niels Fenger (ed), Forvaltningsret (1.udgave, 1. oplag), Jurist- og Økonomforbundets Forlag, 2018, 627-49.

[26] The right to information act (offentlighedsloven) can be found at: https://www.retsinformation.dk/forms/r0710.aspx?id=152299

The Danish Act on Public Administration contains a section on explainability (§§22-24).[27] In general, the explainability requirement can be said to entail that the citizen to whom the decision is directed must be given sufficient information about the grounds of the decision. This means that the explanation must fully cover the decision and not just explain parts of the decision. The explanation must also be truthful and in that sense correctly set forth the grounds that led to the decision. §24 sets out the elements that must be included in an explanation. These are:

1) A reference to those legal rules according to which the decision is made. The reference to legal rules must be clear and specific. It is not sufficient to mention an overall statutory act – the specific articles in the act that provides the legal foundation for the decision must be clear.
2) In so far as the decision involves administrative discretion, the main considerations decisive for the discretionary elements of the decision must be set out. This amounts to a requirement that the main interest driving the decision should be set out explicitly.
3) Information about those factual circumstances that have had a considerable influence on the decision in the case. This part of the requirement should make it clear which facts serve as the basis for the decision

The Danish Parliamentary Ombudsman has supplied an overview of practice in regards to the explainability requirement.[28] It shows that explanations may be limited to stating that some factual requirement in the case is not fulfilled. For example, a certain age has not been reached, a doctor's certificate is not provided or a spouse's acceptance has not been delivered in the correct form. Explanations may also be standard formulations that are used frequently in the same kind of cases. Finally, it does not seem to be possible to formulate any specific standards in regards to how deep or broad an explanation should be in order to fulfil the minimum requirement under the law. The requirement is generally interpreted as meaning that explanations should be truthful and reflect the most important elements of the case that have led to the decision[29].

b. Germany

The general requirement to explain administrative decisions can be found in the Administrative Procedural Code (Verwaltungsverfahrensgesetz, VwVfG) of 1976.[30] The main rationale is the

[27] The full text at: https://www.retsinformation.dk/forms/r0710.aspx?id=161411#Kap6

[28] See: https://www.ombudsmanden.dk/myndighedsguiden/generel_forvaltningsret/begrundelse/

[29] An example may illustrate this. In FOB 2012.17 a couple suffering from cerebral Paresis was seeking artificial insemination. Their application was rejected without specifying why. The Ombudsman criticized lack of explanation, pointing out the disease and its effect on the possibility of caring for a child should have been mentioned. Another concerned a man who was receiving salary compensation from the local municipality because he was ill and therefore incapable of undertaking work. This was documented in a health declaration from the man's doctor. In an application to have prolonged compensation the municipality refused the request, the municipality explained the decision by reference to a renewed health declaration from the doctor who had examined the man. They did not however explain, which part of the new declaration led to their decision to refuse the request. The Ombudsman criticized the explanation for being insufficient. It ought to have specified how the new health declaration differed from the first declaration and how this change was linked to the refusal decision.

[30] Art. 39 VwVfG. Specialised regimes, e.g. for taxes and social welfare, contain similar provisions.

constitutionally protected right to a legal remedy,[31] which would be largely ineffective in the absence of an explanation. As outlined above, the explainability requirement serves also a number of other important functions, including self-control as well as internal and external control of the administration, general democratic acceptance and transparency (e.g. avoiding the impression of secrecy and bias). They find their constitutional protection in the principles of democracy, rule of law and fairness.[32] Generally speaking, every written (or electronic) decision requires an explanation; it should outline the essential factual and legal reasons that gave rise to the decision. In case of discretionary decisions, the explanation should detail the yardstick used in assessing similar cases, and (if relevant) any deviation from such policies.[33] The need for an explanation is even greater for discretionary decisions; they are usually not subject to judicial review.[34] There is therefore wide support in the literature and jurisprudence for the proposition: the wider the margin of discretion of the decision-maker, the more detailed the explanation must be.[35]

It is commonly held in textbooks that administrative decisions that do not adversely affect the citizen in question do not require an explanation.[36] Kischel argues, however, that on many occasions the interests of others (including the general public) and the demands for control and transparency would make an explanation necessary, even in cases of positive decisions.[37]

Another proclaimed exception from the duty to give reasons concerns administrative acts issued in large numbers or with the help of automatic means.[38] However, the added value of the exception is rather limited and has been rightly criticised.[39] Issuing at least standardised explanations for computer-aided decisions (including positive ones) would in no way overburden the administration. Rather, such explanations would come at no additional costs and may lead to better documentation and control as well as greater trust and acceptance of such forms of decision-making among the public. The possibility of so-called "fully automated decisions" was added through a special provision, Art. 35a VwVfG, which entered into force in 2017.[40] It is only a framework provision and requires additional legislation to make use of automated decisions. When doing so, the legislator will also have to comply with the additional requirements under Art. 22 GDPR.[41] Most importantly, Art. 35a excludes the use of fully automated decision-making for cases involving discretionary decisions

[31] Art. 19 (4) GG.

[32] Kischel (n 13) 63–143.

[33] Franz-Joseph Peine and Thorsten Siegel, *Allgemeines Verwaltungsrecht* (CF Müller 2018) 161–162. mn. 517-18. In this regard, Art. 37 VwVfG follows largely the standards set out by the Federal Administrative Court in 1993, BVerwG 1 B 117.83, para. 4.

[34] Kischel (n 13) 223–224.

[35] ibid 224.

[36] Rudolf Schweickhardt, Ute Vondung and Annette Zimmermann-Kreher, *Allgemeines Verwaltungsrecht* (Kohlhammer Verlag 2018) 586.

[37] Kischel (n 13) 229–232.

[38] Art. 39 (2) VwVfG. However, decisions involving individual and specific cases (e.g. asylum procedures) would always require an explanation.

[39] Kischel (n 13) 243–244.

[40] Similar (albeit not identical) provisions exist also in the specialised regimes for taxes and social welfare. See the comparative analysis by Nadja Braun Binder, Weg frei für vollautomatisierte Verwaltungsverfahren in Deutschland, in: Jusletter IT 22 September 2016.

[41] Martini/Nink, Wenn Maschinen entscheiden … – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz, NVwZ – Extra 10/2017, 8.

and a margin of appreciation. According to the Bundestag's commentary to the new provision, such cases continue to require human involvement.[42] This reflects a long-held view among German administrative lawyers sceptical of automated decision-making.[43] Yet, there are good policy reasons against such a strict ban and for keeping the legal framework for automated decision-making open for innovation.[44] In the absence of actual use of fully automated decision-making and related case law, it needs to be seen how the field will develop and what role (if any) explainability can play alongside other measures (including spot-checks, access to information etc.).[45]

c. France

Unlike in Germany and the Scandinavian countries, there is no general explainability requirement for administrative decisions.[46] Indeed, as the Conseil Constitutionnel held in 2004, French constitutional law does not by itself impose a general duty on administrative bodies to explain their decisions.[47] Beyond sanctions of a punitive character, administrative decisions need to be reasoned as provided by a 1979 statute[48] and the 2016 Code des Relations entre le Public et l'Administration (CRPA). The CRPA requires a written explanation that includes an account of the legal and factual considerations underlying the decision.[49] The rationale behind the explainability requirement is to strengthen transparency and trust in the administration, and to allow for its review and challenge before a court of law.[50]Note that those explanations are generally only required for negative decisions ("decisions défavorables").[51] In addition, the law provides public authorities with ample opportunities to invoke the protection of state secrets or other interests (including national defence, foreign policy, public order etc.) so as to avoid an explanation altogether.[52]

Despite its late inclusion of the explainability requirement, France was early in regulating the use of automated decision-making. Indeed, Art. 10 of the 1978 Loi Informatique et Libertés provided for an blanket ban of decisions producing legal effects for individuals while being based solely on automated

[42] BT-Drs. 18/8434, p. 122.

[43] Lazararos, Rechtliche Auswirkungen der Verwaltungsautomation auf das Verwaltungsverfahren (1990) 222-29; B. Degtandi, Die automatisierte Verwaltungsverfügung, (1977) 77-90.

[44] Djeffal, C. Das Internet der Dinge und die öffentliche Verwaltung: Auf dem Weg zum Smart Government? (2017) Deutsches Verwaltungsblatt (DVBl), 808-816, 814-15.

[45] Suggested by Martini/Nink, 10/2017, 14.

[46] J.-L. Autin, La motivation des actes administratifs unilatéraux, entre tradition nationale et évolution des droits européens "RFDA" 2011, no. 137-38, 85-99, 88.

[47] Conseil Constitutionnel 1 juillet 2004, no. 2004–497 DC ("les règles et principes de valeur constitutionnelle n'imposent pas par eux-mêmes aux autorités administratives de motiver leurs décisions dès lors qu'elles ne prononcent pas une sanction ayant le caractère d'une punition").

[48] Loi du 11 juillet 1979 relative à la motivation des actes administratifs et à l'amélioration des relations entre l'administration et le public.

[49] Art. L211-5 ("La motivation exigée par le présent chapitre doit être écrite et comporter l'énoncé des considérations de droit et de fait qui constituent le fondement de la decision").

[50] N. Songolo, La motivation des actes administratifs, 2011, www.village-justice.com/articles/motivation-actes-administratifs,10849.html.

[51] Art. L211-2 ("Les personnes physiques ou morales ont le droit d'être informées sans délai des motifs des décisions administratives individuelles défavorables qui les concernent. A cet effet, doivent être motivées les décisions qui : …"). Note also Art. L211-3 ("Doivent également être motivées les décisions administratives individuelles qui dérogent aux règles générales fixées par la loi ou le règlement").

[52] Art. L211-2 (7) in conjunction with Art. L311-5 (2).

processing of their data. Interestingly, Art. 22 GDPR adopted a similar formulation. Yet, it permits fully automated decisions provided that certain safeguards are in place.[53] Those safeguards have been criticised as largely inadequate.[54] Provided that it does not involve "sensitive data", such procedures can be used for a broad range of administrative decisions.[55] The new Art. 10 from June 2018 requires the sharing of information on the algorithms upon request.[56] However, practice shows a low compliance rate among public authorities. Art. 10 provides that failure to share invalidates the decision, but (after recent amendments to that law) only from July 2020 onwards and only for fully automated decisions.[57]

D. United Kingdom

Though, "there is no general common law requirement for reasons… the common law is recognising a growing number of exceptions to this rule, where reasons are required."[58] In *Doody*, Lord Mustil was quite clear that while "the law does not *at present* recognise a general duty to give reasons for an administrative decision … [it is] broadly beyond question that such a duty may in appropriate circumstances be implied."[59] This emerging common law principle[60] has been reiterated in a number of cases where the duty to give reasons is significant for the person about which a decision is or has been made or for a general principle of fairness in judicial proceedings.[61] Though there is a lack of specific language both considering automatic decisions and a duty to give reasons in general, or a general standard to apply, the requirement of explainability in the form it lives is likely plastic enough to provide significant protection for ADM use.

As Marion Oswald has pointed out, the case law in the UK has a significant history in spelling out what is required when giving reasons for a decision.[62] As she recounts from *Dover District Council*, "the content of [the duty to give reasons] should not in principle turn on differences in the procedures

[53] Art. 10 Loi Informatique et Libertés, 2018,
[54] Élise Untermaier-Kerléo, 'Les nouveaux visagesde la décision administrative : d'une administration assistée à une administration automatisée', 2018, L'administration augmentée, actes de colloque.
[55] Ibid, 2.
[56] Art. 10 (2) in conjuction with Art. L311-3-1 ("Sous réserve de l'application du 2° de l'article L. 311-5, une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l'administration à l'intéressé s'il en fait la demande").
[57] https://www.nextinpact.com/news/106986-obligation-dexplicitation-algorithmes-publics-an-pour-rien.htm.
[58] Andrew Le Sueur, 'Robot Government: Automated Decision-Making and Its Implications for Parliament' (Social Science Research Network 2015) SSRN Scholarly Paper ID 2668201, availble at: https://papers.ssrn.com/abstract=2668201, 9.
[59] *Doody v. Secretary of State for the Home Department* [1993] 3 All E.R. 92 at 110
[60] See for example the "judge over your shoulder (JOYS)" advice on recording reasons as outlined in Carol Harlow and Richard Rawlings, 'Proceduralism and Automation: Challenges to the Values of Administrative Law' in E Fisher, J King and A Young (eds), *The Foundations and Future of Public Law (in honour of Paul Craig)* (OUP Oxford 2019) <https://papers.ssrn.com/abstract=3334783> , 14.
[61] See among others, *R v Higher Education Funding Council*, ex p *Institute of Dental Surgery* [1994] 1 WLR 242;
[62] Marion Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' (2018) 376 *Phil. Trans. R. Soc. A* <https://ssrn.com/abstract=3216435>.

by which it is arrived at."[63] What is paramount in the UK conception is not a differentiation between man and machine but one that stands by enshrined and tested principles of being able to mount a meaningful appeal. As Oswald continues, "administrative law principles governing the way that state actors take decisions via human decision-makers, combined with judicial review actions, evidential processes and the adversarial legal system, are designed to counter […]" any ambiguity in the true reasons behind a decision.[64] An algorithm does not change these safeguards, but they may stretch the boundaries of the considerations taken in a decision beyond their traditional human counterparts. However, the use of a learning algorithm (or even statistical inference) doesn't by necessity violate principle of relevancy in what counts as meaningful in a decision. As long as these inferences are included in the explanation in a legally similar way to a human decision, it would be a hard sell to find an algorithm any more culpable than a human being of obfuscating the 'real' reasons behind a decision. In fact, the opposite may be true given the amount of factors available to challenge on appeal, including a piece of data's relevance as an input factor, how the algorithm has performed in the past, the relevance of factors not included in the model, and the "causal relationships between the inputs and the prediction claimed."[65]

e. EU Law

Art. 41 of the Charter of Fundamental Rights of the European Union (CFR) from 2000 provides for a *right to good administration*, which includes in paragraph 2 the "obligation of the administration to give reasons for its decisions", successfully adopted following proposals from Scandinavian member states.[66] Its inclusion is a concretization of Art. 296 (2) Treaty on the Functioning of the European Union (TFEU), according to which "[l]egal acts shall state the reasons on which they are based", which applies also to administrative decisions.[67] Art. 41 CFR binds primarily EU institutions, but the same rule applies equally to member states implementing EU law.[68] Generally, all unilateral acts that generate legal consequences – and qualify for judicial review under Art. 263 TFEU – require an explanation.[69] It must "contain the considerations of fact and law which determined the decision".[70] There is a clear link between the range of the available discretion and the scope of the duty to give reasons, i.e. decisions need to be "more thoroughly reasoned the greater the discretionary power".[71] The explainability requirement was further concretized by the European Code of Good

[63] *Dover District Council (Appellant) v CPRE Kent (Respondent) CPRE Kent (Respondent) v China Gateway International Limited (Appellant)* [2017] UKSC 79, para 41. See in particular, *Stefan v General Medical Council* [1999] 1 WLR 1293 at page 1300G.
[64] Oswald (n 63) 6.
[65] ibid 14.
[66] Autin, (n 14) 87.
[67] See, https://fra.europa.eu/en/charterpedia/article/41-right-good-administration#group-info-publications.
[68] Herwig CH Hofmann and C Mihaescu, 'The Relation between the Charter's Fundamental Rights and the Unwritten General Principles of EU Law: Good Administration as the Test Case' (2013) 9 European Constitutional Law Review 73, 73–101.
[69] Case C-370/07 Commission of the European Communities v Council of the European Union, 2009, ECR I-08917, recital 42 ("which is justified in particular by the need for the Court to be able to exercise judicial review, must apply to all acts which may be the subject of an action for annulment").
[70] Schwarze, European Administrative Law (Sweet and Maxwell 2006), 1406.
[71] Ibid, p. 1410.

Administrative Behaviour,[72] a soft law document from 2002 aimed at European Commission staff, as well as by EU case-law.[73]

Perhaps the most glaring difference that would arise between automated and non-automated scenarios is the direct application of Art. 22 of the General Data Protection Regulation (GDPR), which applies specifically to "Automated individual decision making, including profiling." Art. 22 stipulates that a data subject "shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her"[74] unless it is proscribed by law with "sufficient safeguards" in place,[75] or by "direct consent."[76] These sufficient safeguards range from transparency in the input phase (informing and getting consent) to the output-explainability phase (review of the decision itself). The GDPR envisages this output phase in the form of external auditing through Data Protection Authorities (DPAs), which have significant down sides in terms of effectiveness and efficiency.[77]

Art. 22 also stipulates that one has the right to contest the decision and to "obtain human intervention" in those situations.[78] There is also a strict prohibition against the use of "special categories" of personal data unless one of the circumstances of Art. 9(2) applies.[79] To make these decisions, an automated system would need access to a repository of data. As to the explainability requirement of such a system, the GDPR specifies "that information is provided" to the data subjects that informs them of the "existence of automated decision making", "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data

[72] European Ombudsman, The European Code of Good Administrative Behaviour, 1 March 2002: Art. 18 ("1. Every decision of the institution which may adversely affect the rights or interests of a private person shall state the grounds on which it is based by indicating clearly the relevant facts and the legal basis of the decision. 2. The official shall avoid making decisions which are based on brief or vague grounds, or which do not contain an individual reasoning passed. 3. If it is not possible, because of the large number of persons concerned by similar decisions, to communicate in detail the grounds of the decision and where standard replies are therefore sent, the official shall subsequently provide the citizen who expressly requests it with an individual reasoning").

[73] A significant recent ruling is *Bamba* (2012), a case involving an asset-freezing measure taken in response to the political crisis in Côte d'Ivoire in 2010-11. Most remarkably, the CJEU set aside the ruling of the General Court, which had previously annulled the decision concerning the applicant on the ground of being insufficiently reasoned. Instead, the CJEU considered the general context of the restrictive measure reasonably well-known to the applicant and thus held that the decision had been in line with the explainability requirements. CJEU, Bamba v Council, Judgment, 15 November 2012, Case C-417 / 11, paras. 49-55. See also: Laura Muzi, 'Administrative due process of law in the light of the jurisprudence of EU Courts: a quantitative and qualitative analysis' in: Carol Harlow, Päivi Leino, Giacinto della Cananea (eds.), *Research Handbook on EU Administrative Law* (Elgar 2017), 468-89.

[74] Reg (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Dir 95/46/EC (General Data Protection Regulation) 2016, Art. 22(1).

[75] ibid, Art. 22(2)b.

[76] ibid, Art. 22(2)c.

[77] See Antoni Roig, 'Safeguards for the Right Not to Be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)' (2017) 8(3) *European Journal of Law and Technology*.

[78] GDPR, Art. 22(3).

[79] ibid, Art. 22(4).

subject."[80] What the make-up of *the logic* and *meaningful* are is far from clear.[81] There is no consensus on whether this means a logic connoting *how* (the components of the decision making process) or *why* (the outcome of those components) a decision was made. Even recital 71, which includes a right "to obtain an explanation of the decision reached after such assessment",[82] is not of particular help in this regard.

One system that may help us here is the interpretation surrounding the legal requirements of administrative fairness in the European Convention of Human Rights.[83]

f. European Convention on Human Rights

While there is no direct reference to explainability in the ECHR in the clearest terms as it is in the GDPR or national laws, it is worth considering the protections afforded to individuals regarding administrative decision-making (both automated and human). We address explainability through the Convention's protection of redress and safeguards regarding a 'fair trial' under article 6 as interpreted by the European Court of Human Rights (ECtHR).[84] The first question is whether our scenario would fall under the remit of article 6. Administrative decisions, such as the payment of loss earnings, are well enshrined in the Court's jurisprudence as 'public law'[85] including those that include a right to the administrative documents[86] and put them well within the Court's jurisdiction. However, the Court has carved out exceptions to this, such as tax proceedings[87], immigration[88], and the granting of

[80] ibid, Art 13(2)f, Art 14(2)g, Art 15(1)h.

[81] Bryce Goodman and Seth Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2017) 38 AI Magazine 50. *Cf.* Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 76.

[82] GDPR, Recital 71. Recital 71 reiterates the right not to be subject to a decision "which may include" an automated aspect. However, its threshold and assessment criteria do not do much by way of clarification. It leaves the "fair and transparent processing" to be determined by "appropriate" "mathematical or statistical procedures", "organisational measures" where inaccuracies are corrected and "risk of errors is minimised". What is appropriate or significantly or reasonably minimised, again, is hard to surmise. Given the dearth of case law available on Art. 22, it is hard to know in our scenario what exactly is required to be explained at either the initial decision or appeal level.

[83] It should be noted, that there may be a tentative avenue to explore explainability under the 'right to information' under articles 2, 8, & 10. However, for brevity, that will not be discussed in this paper. These are not the only criteria of fairness (there are also questions of receiving a hearing in one's presence, effective participation, presumption of innocence, freedom from self-incrimination, and principle of immediacy, and impartiality), but equality of arms and a reasoned judgement are sufficient for examining the legal requirement of explainability.

[84] Though there are many aspects to consider with regard to article 6, we focus mainly on the requirement of equality of arms as it applies to explainability.

[85] See among many others: *Sporrong and Lönnroth v Sweden* [1982] ECHR 7152/75, ECLI:CE:ECHR:1982:0923JUD000715175, §79; *Benthem v The Netherlands* [1985] ECHR 8848/80, ECLI:CE:ECHR:1985:1023JUD000884880, §36; *Tre Traktörer Aktiebolag v Sweden* [1989] ECHR 10873/84, ECLI:CE:ECHR:1989:0707JUD001087384, §43; *Tinnelly & Sons Ltd and Others and McElduff and Others v the United Kingdom* [1998] ECHR 20390/92, ECLI:CE:ECHR:1998:0710JUD002039092, §61.

[86] *Loiseau v France* [2004] ECHR 46809/99, ECLI:CE:ECHR:2004:0928JUD004680999.

[87] *Ferrazzini v Italy* [2001] ECHR [GC] 44759/98, ECLI:CE:ECHR:2001:0712JUD004475998, §25.

[88] *Maaouia v France* [2000] ECHR [GC] 39652/98, ECLI:CE:ECHR:2000:1005JUD003965298, §38.

passports,[89] among others based on the "part of the hard core of public-authority prerogatives"[90] not considered under Art. 6. Given the nature of our scenario, a decision regarding social welfare payments is a closer analogue to the scenarios which are "decisive for private rights and obligations"[91] in cases like *Feldbrugge* (dealing with a medical appeals board)[92] than a tax case dealing with the transfer of property in *Ferrazzini*.[93] It is safe to assume that even a fully automated system that dealt with the payments of social welfare benefits would be under the remit of Art 6 considering the numerous cases dealing with welfare including invalidity,[94] disability[95], and housing, among various others. That being said, whether all of the models of decision-making outlined above (pure, hybrid, or automated) would count as a *tribunal* for the purposes of Art. 6 would be highly context-dependent. The conventional wisdom on the determination of a 'tribunal' comes from *Benthem*, where the Court stated that "a power of decision is inherent in the very notion of 'tribunal' and that a mere advisory role (the case with ADM support models) would not suffice.[96] All three models would be considered a tribunal for Art. 6, assuming they had a determining binding force vs a purely advisory role and that the decision is final. In other words, the machines would be considered the same.

Next, we can ask what would change with automated processing between human and machine with regard to the fairness of the proceedings themselves, particularly in terms of the equality of arms and right to a reasoned judgement, between an applicant and an algorithmic assessor? The equality of arms principle states that, "each party to be given a reasonable opportunity to present his case under conditions that do not place him at a substantial disadvantage."[97] With regard to our scenario, the case of *Hentrich v. France* is particularly illustrative of the Court's approach to equality of arms in administrative decisions. There the Court considered that the fairness requirement was not met given that the reasons given for the administrative decision in that case were "too summary and general to enable Mrs Hentrich to mount a reasoned challenge to that assessment."[98] Though there may not be any explicit right to an explanation for decisions, the equality of arms principle is clear that the fairness of proceedings must require an explanation to the appealing party that is specific enough to mount a reasoned challenge. This reasoned challenge would include understanding both the "facts and procedures in which factual findings…were arrived at," to assess and/or challenge whether

[89] *Sergey Smirnov v Russia* [2009] ECHR 14085/04, ECLI:CE:ECHR:2009:1222JUD001408504.
[90] Council of Europe, 'Guide on Article 6 of the European Convention on Human Rights Right to a Fair Trial (Civil Limb)' (2018) <https://www.echr.coe.int/Documents/Guide_Art_6_ENG.pdf>.at §65
[91] ibid. at §30
[92] *Feldbrugge v the Netherlands* [1986] ECHR 8562/79, ECLI:CE:ECHR:1986:0529JUD000856279; see also, *Deumeland v Germany* [1986] ECHR 9384/81, ECLI:CE:ECHR:1986:0529JUD000938481.
[93] *Ferrazzini* (n 88).
[94] *Schuler-Zraggen v Switzerland* [1993] ECHR 14518/89, ECLI:CE:ECHR:1993:0624JUD001451889.
[95] *McGinley and Egan v UK* [1998] ECHR 21825/93 and 23414/94, ECLI:CE:ECHR:1998:0609JUD002182593; *Salesi v Italy* [1993] ECR [GC] 13023/87, ECLI:CE:ECHR:1993:0226JUD001302387; *Tsfayo v the United Kingdom* [2006] ECHR 60860/00, ECLI:CE:ECHR:2006:1114JUD006086000.
[96] *Benthem* (n 86), §40.
[97] *Kress v France* [2001] ECHR [GC] 39594/98, ECLI:CE:ECHR:2001:0607JUD003959498, §72.
[98] *Hentrich v France* [1994] ECHR 13616/88, ECLI:CE:ECHR:1994:0922JUD001361688, §56.

certain considerations were true, relevant, and rational, but not to the point where it requires a full reopening of the case.[99]

g. What Explainability Requires

Explainability is not different now that it is algorithmic. The different approaches outlined here provide significant thresholds that any ADM system would need to pass without requiring the full opening of a black box. In each of the compared legal regimes there is a mix of protections that construct a threshold that is robust enough to cover the majority of models of ADM. In our example, we might assume for instance that Parent A is challenging a negative decision that the illness in question was considered serious, chronic or long term. In the above human rights example, in a *pure model* the decision maker would be required to explain the relative facts and procedures, such as the lack of e.g. a doctor's note describing the illness as such, a lack of similar diagnoses being labelled as such, and any particular weighting system applied to that decision, etc. An approach from national and EU law would generally require a detailed explanation outlining the essential factual and legal reasons that gave rise to the decision. The greater the discretionary power of the *pure* decision maker, the more thorough the explanation has to be. In fact, it should detail the yardstick used in assessing similar cases, and (if relevant) any deviation from such policies. To safeguard the interests of others and ensure better self-control and transparency, those standards may even apply to positive decisions.

A hybrid system that was rule-based or similarly transparent would not require any different meaningful explanations. The fear is the fully automated, black box system that would be impossible to give an answer of this sort. However, this is not the case. Requirements of being informed and strict consent derived from both national laws and EU law would allow Parent A to at least know the information that is being considered in making those decisions. The fear then is not knowing what parameters were weighted and being able to test for any type of bias to an extent that Parent A would be as comfortable as in the purely human scenario. For this, a rewrite of legal requirements is not necessary. Instead, we propose what we have labelled an 'administrative Turing test'.

5. Ensuring Legal Quality through Hybrid Systems

Introducing a machine-learning algorithm in public administration and using it to produce drafts of decisions may advance efficiency in case administration and decision-making without lowering legal quality as long as the data the algorithm is learning from is sufficiently large and generally contains correct and well-reasoned legal decisions. Learning from historical cases and reproducing their language in new cases by connecting legal outcomes to given fact descriptions is not far from what human caseworkers would do anyway: Whenever a caseworker is attending to a new case, he or she will seek out former cases of the same kind to use as a compass to indicate how the new case should be decided. The difference between the human and the algorithm is that algorithms tend to be more rigorous than humans. Humans respond more organically to past cases because they have a broader horizon of understanding: They contextualize their task to a much richer extent, and can therefore

[99] See, *Fazia Ali v the United Kingdom* [2015] ECHR 40378/10, ECLI:CE:ECHR:2015:1020JUD004037810, §83.

adjust their decisions to a broader spectrum of facts – including ones that are hidden from the explicit law (e.g. resource allocation and policy). It is precisely this phenomenon that can explain why new practice can develop under the same law.[100]. Algorithms on the other hand operate without such context and can only relate to explicit texts. Hence they cannot evolve in the same way. Paradoxically then, having humans in the legal loop serves the purpose of relativizing strict rule-following.

This limited contextualization of algorithmic "reasoning" will create a problem if *all* new decisions are drafted on the basis of an algorithm that reproduces the past and if those drafts are only subjected to minor or no changes by its human collaborator. The reason is the following: once the initial learning stage is finalized and the algorithm is used in output mode to produce drafts, then new decisions will be based on drafts produced by the algorithm. One of two different situations may now occur: one, the new decisions are fed back into the machine-learning stage. In this case, a feedback loop is created in which the algorithm is fed its own decisions. Or two, the machine-learning stage is blocked after the initial training phase. In this case, every new decision is based on what the algorithm picked up from the original training set. None of these options are in our opinion optimal for maintaining an up-to date algorithmic support system.

There are good reasons to think that an algorithm will only keep performing well (which in this case is measured by the algorithm's ability to issue usable drafts of a good legal quality) – if it is constantly maintained by fresh input[101]. This can be done in a number of different ways, depending on how the algorithmic support system is implemented in the overall organization of the administrative body and its procedures for issuing decisions. As mentioned previously, our focus is on models that engage AI and human collaboration. We shall here propose two such principles for organizing algorithmic support in an administrative system that aims at issuing decisions, that we think is particularly helpful because it simultaneously enhances public trust and thereby justification for the continued use of algorithmic decision-support.

In our first proposed model, the case load in an administrative field that is supported by algorithmic decision assistance is randomly split in two loads, such that one load (e.g. 80%) is fed to the algorithm for drafting and another load (e.g. 20%) is fed to a human case worker, also for drafting. Drafts are subsequently sent to a caseworker, who finalizes and signs off on the decisions. All final decisions are pooled and used to regularly update the algorithm used.

By having human administrators interact with algorithmic drafting in this way, and feeding decisions, all touched by a human hand, back into the machine-learning process, the algorithm will be kept "fresh" with new original decisions, a percentage of which will be written by humans from scratch. The effect of splitting the case load and leaving one part to through a "human only" track is that the above mentioned sensitivity to broader contextualization is fed back into the algorithm and hence allows a development in the case law that could otherwise not happen. Although human decision-making is also build from routine and former practice – that, after all is the *raison d'être* of bureaucratization – by singling out a part of the case load to be manually handled and making the

[100] See also Harlow and Rawlings op. cit. who note (at p. 6 in the SSRN version) that: "Administrative Law cannot be static, and the list of values is not immutable; it varies in different legal orders and over time".
[101] See the discussion of the problem with feedback loops in O'Neil (n 2), among others.

human caseworkers aware of the overall working of the system, could well heighten their attention to their role in assuring that decisions are up to "present day conditions" (to paraphrase the ECtHR).

Furthermore, if drafting is kept anonymous, and all final decisions are signed off by a human, recipients of decisions (like our Parent A) will not know or have access to how the decision was produced. Still the explanation requirement assures that recipients can at any time challenge the decision, by inquiring further into the legal justification. What recipients cannot do, however, is demand insight into the underlying neurological or algorithmic computations of caseworkers (human or robotic) – an insight which we have argued above is not legally relevant. We think this way of introducing algorithmic support for administrative decisions advances many of the efficiency gains sought by introducing algorithmic support systems, while preserving the legal quality of decisions.

An alternative method – our second proposed model - is to build into the administrative system itself a kind of continuous administrative Turing test. Alan Turing, in a paper written in 1950[102], sought to identify a test for artificial intelligence (in the paper, Turing asked the question: "Can machines think?"). The test he devised consisted of a set up in which (roughly explained) two computers were installed in separate rooms. One computer was operated by a person – the other was operated by an artificial intelligence system (a machine). In a third room, a human judge was sitting with a third computer. The judge would type questions on his computer and the questions would then be sent to both the human and the AI in the two other rooms for them to read. They would then in turn write replies and send those back to the judge. If the judge could not identify which answers came from the person and which came from the AI (the machine), then the AI would be said to have shown ability to think.

Akin to this, an administrative body could implement algorithmic decision support in a way that would imitate the set-up described by Turing. This could be done in the following way: A certain percentage of the entire case load – say 10% – could be given both to a human caseworker and to an algorithm. Both the human caseworker and the algorithm would produce a decision draft for the same case. Both drafts would be sent to a human judge (i.e. a caseworker who finalizes and signs off on the decision). In this set-up, the human judge would not know which draft came from the algorithm and which came from the caseworker (and formats for issuing drafts could be formalized so as to reduce the possibility of guessing merely by recognizing the style of the drafter's language), but would simply proceed to finalize the decision based on which draft was most convincing for deciding the case and providing a satisfactory explanation to the citizen (Parent A). This final decision would then be fed back to the machine-learning algorithm – for fresh learning.

The two methods described above are both hybrid models and can be used either alone or in combination to assure AI is implemented in a way that is both productive, because drafting is usually a very time consuming process and safe (even if not mathematically transparent) because there is a human overseeing the final product and a continuous human feedback to the drafting system. Moreover, using this hybrid approach helps overcome the legal challenges that a fully automated

[102] A. M. Turing, Computing Machinery and Intelligence. (1950) *Mind* 49: 433-460.

system would face from both EU-law (GDPR) and some domestic legislation (see e.g. Germany above).

Relying on the above models keeps a "human in the loop" and does so in a way that is systematic and meaningful because our models take a specific form: they are built around the idea of human-AI collaboration. Relying on this model makes it possible for tech-companies to develop AI systems that can be introduced to enhance the effectiveness and quality of public administration. The advantage of this is that AI can be developed in a legal environment and be adapted to this. Such an approach, we think, will be optimal for providing working conditions in which AI, in the long term perspective, can grow into a means for assuring better detection of hidden biases and other bureaucratic deficiencies. This approach may help allay the fears of the black box. In terms of control and responsibility, the administrative Turing test allows for a greater scope of review of rubber stamp occurrences by being able to compare differences in pure human and pure machine decisions by a human arbiter (or statistical modelling). With reference to human dignity, the hybrid model retains the human standard as the standard for decision-making. Lastly, full transparency as causal or mathematical explanation does not assuage the fear and consequences of bad data. However, legal thresholds of explanation between fact and legal arguments, as required by fairness in proceedings and a duty to give reasons can expose bias at a greater rate than purely human models. Our Turing model also continually adds new context into the system, allowing for a legal transparency that can protect against ADM models' worst implementations. It goes beyond the general ethical guidelines to impose the law on the books and emphasise its strengths as an enshrined and agreed upon principle that can do the heavy lifting for enforcement. Applying the test developed in this paper today is the most efficient way of overcoming the weaknesses of purely human decision-making tomorrow.

Author(s): Henrik Palmer Olsen, Jacob Livingston Slosser, Thomas Troels Hildebrandt, Cornelius Wiesener
Title: What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration
iCourts Working Paper, No. 162, 2019

Publication date: 12/06/2019

URL: http://jura.ku.dk/icourts/working-papers/

iCourts Working Paper Series
ISSN: 2246-4891

Henrik Palmer Olsen, Professor of Jurisprudence, Faculty of Law, Danish National Research Foundation's Centre of Excellence for International Courts (iCourts), University of Copenhagen; henrik.palmer.olsen@jur.ku.dk

Jacob Livingston Slosser, Carlsberg Postdoctoral Research Fellow, Faculty of Law, Danish National Research Foundation's Centre of Excellence for International Courts (iCourts), University of Copenhagen; jacob.slosser@jur.ku.dk

Thomas Troels Hildebrandt, Professor in Software Engineering, Software, Data, People & Society Section, Department of Computer Science, University of Copenhagen, Denmark; hilde@di.ku.dk

Cornelius Wiesener, Postdoctoral Research Fellow, Faculty of Law, Danish National Research Foundation's Centre of Excellence for International Courts (iCourts), University of Copenhagen; cornelius.wiesener@jur.ku.dk

The iCourts Online Working Paper Series publishes pre-print manuscripts on international courts, their role in a globalising legal order, and their impact on politics and society and takes an explicit interdisciplinary perspective.

Papers are available at http://jura.ku.dk/icourts/

iCourts
- The Danish National Research Foundation's Centre of Excellence for International Courts
The Faculty of Law
University of Copenhagen
Karen Blixens Plads 16
2300 Copenhagen S
E-mail: icourts@jur.ku.dk
Tel. +45 35 32 26 26